The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Improving feature based object recognition in service robotics by disparity map based segmentation.

D. Asanza and B. Wirnitzer

Abstract— **This article presents a novel object recognition module which is adapted to the needs in mobile service robotics. It uses information provided from a stereo camera system as pre-processing part of SIFT or SURF. The principle idea is to filter irrelevant information by selecting regions of interest in the disparity map from stereo images and to use the geometrical constraints of the stereo camera system in order to filter out useless descriptors in early stages of the processing chain. Experimental results show that this setup improves overall performance in comparison to similar systems by a factor of two to five.**

I. Introduction

Robust and fast object recognition is an important component in mobile service robotics, especially when the robots will have to deal with everyday life environments [1], [2]. In actual service robot projects object recognition is often based on the SIFT or SURF algorithm, proposed by Lowe [3] or Bay et. al. [4], [5]. Though the details of both algorithms differ, SIFT and SURF have three steps in common (Fig. 1): i) detection of interesting points, also called keypoints, ii) formation of scale and rotational invariant descriptors at each keypoint and iii) feature matching to find the searched object.

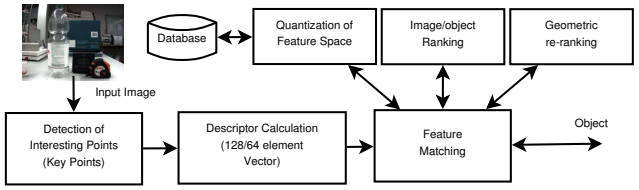

Fig. 1. Processing modules of feature based object recognition based on standard SIFT and SURF

Detection of interesting points in the SIFT algorithm is achieved by searching for minima and maxima in the Difference of Gaussian (DoG). Down-sampling of Gaussian filtered images is used for scale invariant detection. In SURF interesting point selection is based on an approximately calculated Hessian-matrix. The trace of the Hessian, which corresponds to the Laplace operator, is used to select the scale and the determinant of the Hessian to select the location. The great advantage of the Hessian approach is that it can be approximated using very fast box filters; its draw back is the slightly reduced repeatability under image rotations around ± 45 degrees compared to the DoG approach.

In SIFT scale and rotational invariant descriptors are formed by the histogram of local oriented gradients around the interest point and stored in a 128-dimensional vector. SURF descriptors are 64-dimensional vectors build on the distribution of first order Haar wavelet responses in x and y direction.

Feature matching based on the Mahalanobis or Euclidean distance is the final step in the object recognition. In the training phase the SIFT/SURF features of objects are collected and stored. In the recognition phase descriptors are calculated for the whole image and then compared with the stored descriptors. An object is recognized if at least some of its descriptors match the stored descriptors for that object.

In the case of a dataset containing millions of objects feature matching is a difficult task with high computational costs. In order to reduce the storage requirements and to speed up the search recent research has concentrated to improve this step by i) quantization of the feature space, ii) image or object ranking and iii) geometric re-ranking.

Quantization of the feature space and image or object ranking can be achieved according to Sivic and Zisserman [5]: local image descriptors are quantized and clustered into "visual words". Quantized matching is performed using a bag-of-words (BOW) method in which visual word occurrences alone are used to measure image similarity. Their approach employs a term-frequency inverse document-frequency (tf-idf) weighting scheme similar to that used in text retrieval. In the context of SIFT and SURF similar descriptors in an image are searched, quantized into visual words and discarded while their geometric data are preserved for later matching. This can provide significant savings in memory compared to storing and matching each feature independently. This and similar approaches have been developed further [6],[7],[8],[9],[10] and have been used especially to recognize buildings in data sets containing millions of images. In [7] a system is described which uses a vocabulary tree based matching algorithm and which can search a database with more than 40 000 images in a few milliseconds. These different systems have in common that they end up in a ranked list of probable objects in the data base which to some degree match the actual object under investigation.

Geometric re-ranking is the final step in the matching procedure. A popular method is for example to use RANSAC to extract those features which fit best to a known geometrical

This work was supported by Zentrum für angewandte Forschung an Fachhochschulen *http://www.zafh-servicerobotik.de*

D. Asanza is Research Assistant at the Laboratory for Digital Signal Processing. Mannheim University of Applied Sciences. Paul-Wittsack-Straße 10, 68163, Mannheim, Germany. asanza@hs-mannheim.de

B. Wirnitzer is head of the Institute for Digital Signal Processing, Mannheim University of Applied Sciences. Paul-Wittsack-Straße 10, 68163, Mannheim, Germany. b.wirnitzer@hs-mannheim.de

model of the object.

This paper presents a novel object recognition module which is adapted to the needs in mobile service robotics. It uses information provided from a stereo camera system as pre-processing part of SIFT or SURF. The principle idea is to use the disparity map from stereo images and the geometrical constraints of the stereo system in an early stage of the processing chain in order to filter irrelevant information and useless descriptors. Surprisingly a very simple, not very perfect but fast method for selecting regions of interest in the scene which uses connected component analysis on a disparity map, was sufficient to improve performance parameters of our object recognition system by a factor of two to five. Section 2 outlines the concept of the system and its components: the stereo cameras, the disparity based object segmentation and the epipolar filtering. In Section 3 we present experimental results that outline the improvements achieved in everyday life scenarios.

II. Concept of the System

The model of our improved object recognition system is depicted in Fig. 2. The stereo cameras provide valuable information that helps to reduce the data in early stage of the processing chain: i) Object segmentation is achieved by connected component analysis on the disparity map calculated between the right and the left images. This reduces the amount of data transferred to the keypoint detector. ii) The epipolar constraint, which is a geometric property of the stereo camera system, is used to filter all detected key points which are not consistent with the stereo camera model. iii) Finally, only those descriptors which have a correspondence in the right and in the left image are transmitted to the feature matching module. The detailed description of the system modules is given in the following subsections.

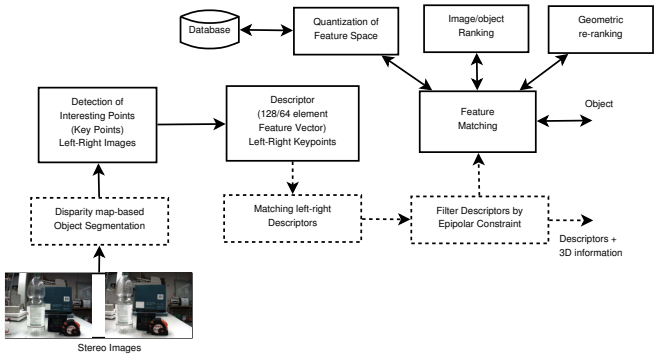

Fig. 2. Proposed System

A. Stereo Cameras

Figure 3 illustrates a stereo camera system. Here, a point $P = [X, Y, Z]$ in the space is projected into a point in the image planes. The homogeneous coordinates of the projected points $x_l = [x, y, w]$ and $x_r = [x', y', w']$ are determined by the intrinsic matrix of each camera, as shown in equation 1. The intrinsic matrix is a 3×3 matrix which encodes the focal distance and coordinates of the center of the image planes. Parameters f,c_x, c_y are called "Intrinsic Parameters".

$$\begin{bmatrix} x \\ y \\ w \end{bmatrix} = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{1}$$

The equation above corresponds to an ideal camera. In reality, the lenses used and the construction process of the camera introduce distortions which are mainly of two kinds. Radial distortion which occurs because in spherical lenses rays farther from the center of the lens are bent more than those closer in, and tangential distortion which is due to manufacturing defects resulting from the lens not being exactly parallel to the imaging plane. These distortions can be modelled as follows[11]:

Radial Distortion:

$$\begin{bmatrix} x \\ y \end{bmatrix}_{corrected} = \begin{bmatrix} x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \\ y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \end{bmatrix} \tag{2}$$

Tangential Distortion:

$$\begin{bmatrix} x \\ y \end{bmatrix}_{corrected} = \begin{bmatrix} x + 2p_1 y + p_2(r^2 + 2x^2) \\ y + p_1(r^2 + 2y^2) + 2p_2 x \end{bmatrix} \tag{3}$$

To determine the intrinsic parameters (f, c_x, c_y), the radial parameters (k_1, k_2, k_3) and the tangential parameters (p_1, p_2) equations 1, 2 and 3 are solved for know values of x, y, $x_{corrected}$ and $y_{corrected}$ that are obtained from images of a known pattern, in a procedure called camera resectioning or camera calibration. Additionally, the calibration procedure determines the rotation matrix and translation vector that relates the right camera coordinates with that of the left camera.

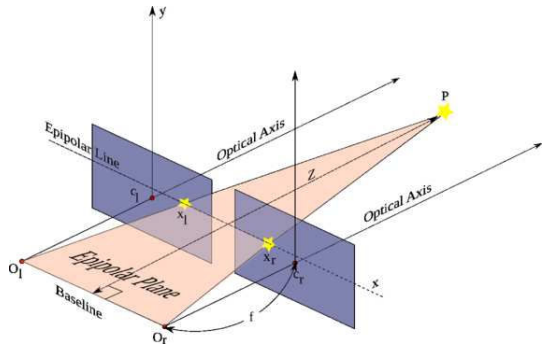

Fig. 3. Stereo Camera System

The advantage of a stereo camera is that it is possible to determine the location of an object in the 3D space using triangulation. To do this, it is important to find the projection of a point in the space in the image planes. The difference between the coordinates of the projected points x_l and x_r is called disparity from which the depth can be calculated as follows:

$$depth = \frac{Baseline \times focal\ distance}{disparity} \tag{4}$$

Using the depth information that corresponds to the Z coordinate of the point in the 3D space, the X and Y coordinates can be calculated from the x and y coordinates in the plane as follows:

$$Y = \frac{(y - c_y)Z}{f} \tag{5}$$

$$X = \frac{(x - c_x)Z}{f} \tag{6}$$

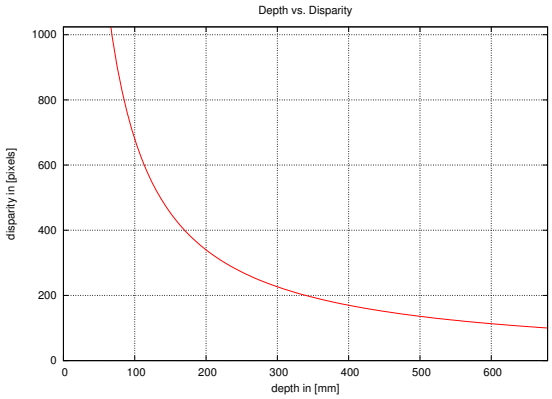

Fig. 4. Meassured depth vs Disparity. Baseline = 5mm, Focal distance = 4mm, Size of Pixel = 4.65μm

Thus, in order to calculate the disparity it is necessary to solve the correspondence problem, which is to find which point in the left image corresponds to which point in the right image. That is a hard problem to solve and no perfect procedure yet exists. However, good results are obtained using a block matching, where small blocks of one of the images serve as pattern to match in the right image. The system shown in figure 2 uses SAD (Sum of absolute differences), as proposed in [12].

Fig. 5. A stereo image pair

To find the correspondent points it is not necessary to search the whole image, but only along a single line. Indeed, one of the geometric properties of a stereo camera system is that correspondent points must satisfy the following relation:

$$x_l^T \times E \times x_r = 0 \tag{7}$$

which is called "Epipolar Constraint". Here x_l and x_r are correspondent points in the left and right images respectively, while E is a 3×3 matrix called the "Essential Matrix". This matrix is obtained from the rotation matrix and translation vector that relates the left camera coordinates with the right camera coordinates, as shown in equation 8.

$$E = R \times S \tag{8}$$

R is the rotation matrix and S is constructed from the Translation vector $T = [T_x, T_y, T_z]$ as follows [13]:

$$S = \begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{bmatrix} \tag{9}$$

B. Object Segmentation

If the values of disparity from each point in the images are organized as a two dimensional matrix the result is a "disparity map". As shown in figure 6 a disparity map is a grayscale image where the pixel values represent the distance from the objects to the camera. Here, black pixels represent points in the scene where no correspondent points could be found. It occurs if not enough texture for block matching is in the scene or because of occlusions between objects that are one in front of the other.

Fig. 6. Disparity Map from the stereo pair of figure 5

An easy way (although not a good one) to segment objects in the image, is to look for connected components in the disparity map. Since occlusions occurs when two objects are one in front of the other, and these occlusions correspond to black pixels around the object in the disparity map, a simple search for connected components can –in most of the cases– segment objects in the image. To enhance the search for connected components, morphological opening is used to widen the gap around the objects [14], as shown in figure 7. A fixed threshold can be applied in order to filter out objects that are too far away to be of interest.

C. Filtering the Features

Connected components found in previous steps are labeled and used as masks to select regions of interest in the left and right images (Fig. 8). In these regions the probability to find an object is high. In order to identify them, SURF features

Fig. 7. Effects of morphological opening with a 7×7 structuring element on the disparity map shown in figure 6

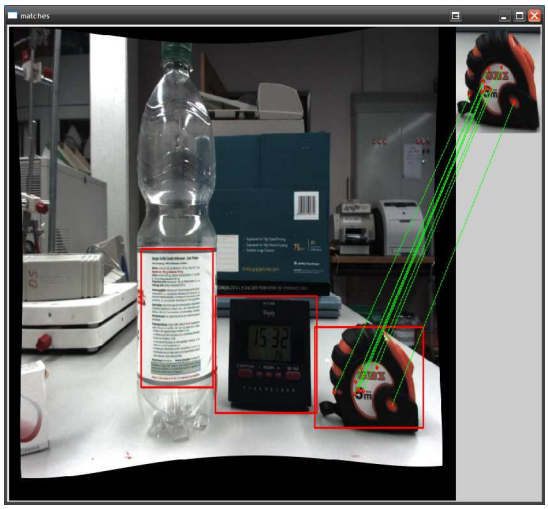

Fig. 9. Example of a matched features. Red rectangles are regions of interest selected in the disparity map.

are extracted from these regions in the left and right images. To filter bad features a two step procedure is used. In the first step features in the left image with no matching in the right image are deleted, while in the second step features that do not satisfy the epipolar constraint from equation 7 are eliminated.

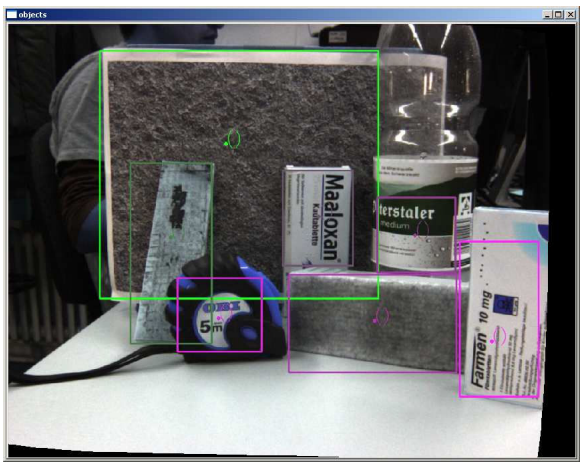

Fig. 8. Segmented objects

The filtered features are then used to identify the object (Fig. 9). Furthermore, as these features are from stereo images, it is also possible to determine their location in the 3D scene.

III. Experiments

In this section we experimentally verify the improvements achieved by the proposed system which uses information provided by a stereo camera as pre-processing part of SURF. We compare a i) standard SURF system as illustrated in Fig. 1, ii) the same system plus filtering of the keypoints by the epipolar constraint (E+SURF) and iii) the same as ii) plus the use of disparity map based object segmentation in 3D space. (D+SURF Fig. 2).

The evaluation of these systems is performed in the context of recognition of the same object observed under different viewing conditions, i.e. different ilumination and location of the object in the scene. The test dataset contains three images of different objects and twenty stereo images of these objects in an unknown environment, with different possitions and orientations. The systems are compared using the same evaluation scenario and test data.

The evaluation criterion is that used in [15]. It is based on the number of correct and false matches obtained between the image of the object and the image of the scene. Since we use images from a real scene, it is difficult to verify the number of correct and false matches because there is no linear transformation relating object and scene images[15]. To overcome this limitation we use planar test objects and prepare the ground truth by manually selecting points to calculate a homography between the object and the scene.

Descriptor matching is performed using a nearest-neighbour ratio matching strategy as described in [4]. The Euclidean distance between each descriptor vector of the keypoints found in the object image and those found in the scene image is calculated. A matching pair is detected if its distance is closer than t times the distance of the second nearest neighbour [4]. The results are presented on the figure 10 as a recall vs. precision curve for values of t varying between 0 and 1. As stated in [16] recall and precision are calculated as follows:

$$Recall = \frac{True\ Positives}{Total\ Positives} \tag{10}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{11}$$

Recall is then a measure of the number of relevant (true) features found by the match procedure, but it says nothing about the irrelevant features found. Precision, on the other hand, show how many irrelevant (false) features where found

for the same match algorithm. Values in Fig. 10 are the average for the dataset.

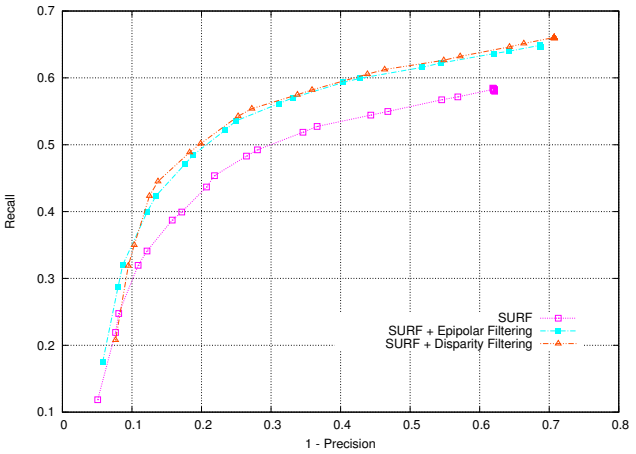

Fig. 10. Evaluation of descriptors obtained with the methods described in this paper. Values calculated for a Nearest Neighbour matching strategy with a threshold value between 0 and 1. (DSURF 31 correspondences, ESURF 33 correspondences, SURF 42 correspondences)

As shown in the figure 10, the use of the epipolar constraint to filter image features has the effect to decrease the number of irrelevant features, thus increasing the recall rate for the same precision, while as shown in Fig. 11 it increases the time needed to search for features by a factor of two, that because two images of the same scene are analysed. On the other hand, the segmentation of the image into regions of interest reduces the time needed to search for features by a factor dependent of the scene. Both, epipolar filtering and selection of regions of interest in the scene, reduce the number of features to analyse, but while the epipolar constraint filter out mostly false features, segmentation of the image eliminates features that are not of interest for the identification of the object, as for example features from objects in the background.

The time of execution shown in figure 11 was measured on an AMD Athlon Dual Core processor computer with 2 GB of RAM and a Linux SMP $x86\_64$ kernel.

IV. Conclusions and Future Work

As shown in section III selection of regions of interest, i.e. reducing the size of the scene where features are searched for, improves the performance of feature-based object recognition systems. Additional information from stereo imaging provides a reliable and easy way not only to select regions of interest but also to filter out useless features in the scene. Surprisingly, though the disparity calculation of the stereo images by itself is slow, the overall performance of the system showed increased number of relevant features retrieved for the same precision compared to standard surf (Fig. 10), while the processing time needed to search and match descriptors showed an improvement by a factor of 3 on average (Fig. 11), although it should be noted that this reduction is highgly dependent on the scene. Future work will focus on better segmentation strategies and the use of

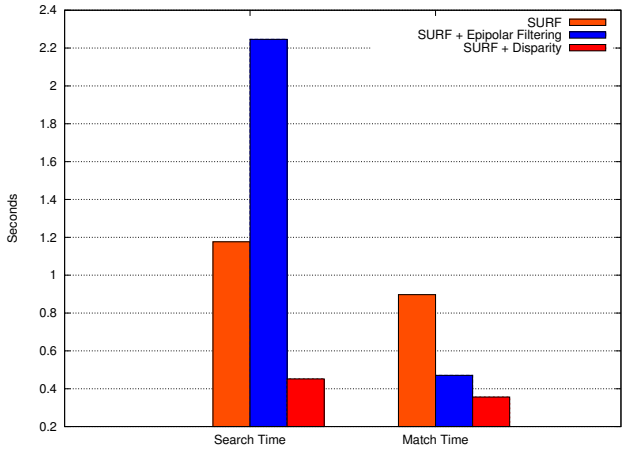

Fig. 11. Comparison of feature search time and descriptor match time.

geometric information in order to improve the recognition rate.

References

[1] J. Kuele, T. Grundmann, and A. Verl, "6d object localization and obstacle detection for collision-free manipulation with a mobile service robot." *Proceedings of the 14th International Conference on Advanced Robotics (ICAR), 22-26 Jun. 2009, Mnchen, Germany*, 2009.

[2] S. Hochdorfer, M. Lutz, and C. Schlegel, "Lifelong localization of a mobile service-robot in everiday indoor environment using omnidirecccional vision." *Proc. IEEE. Int. Conf. on Technologies for Practical Robot Applicatons (TePRA), Woburn, Massachusetts, USA.*, 2009.

[3] D. Lowe, "Distinctive image features from scale-invariant keypoints." *International Journal of Computer Vision*, pp. 91–110, 2004.

[4] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding (CVIU).*, vol. 110, no. 3, pp. 346 – 359, 2008.

[5] J. Sivic and A. Zimmerman, "Video google: A text retrieval approach to object matching in videos." *International Conference on Computer Vision*, pp. 1470 – 1477, 2003.

[6] H. J. M. Douze and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search." *European Conference on Computer Vision.*, pp. 304 – 317, 2008.

[7] N. David and S. Henrik, "Scalable recognition with a vocabulary tree," *IEEE Conference for computer vision and pattern recognition*, 2006.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman., "Object retrieval with large vocabularies and fast spatial matching," *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[9] Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman., "Lost in quantization: Improving particular object retrieval in large scale image databases." *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[10] P. Turcot and D. G. Lowe., "Better matching with fewer features: The selection of useful features in large database recognition problems." *CCV Workshop on Emergent Issues in Large Amounts of Visual Data (WS-LAVD), Kyoto, Japan.*, 2009.

[11] J. G. Fryer and D. C. Brown, "Lens distortion for close-range photogrammetry," *Photogrammetric Engineering and Remote Sensing*, vol. 52, pp. 51–58, 1986.

[12] K. Konolige, "Small vision system: Hardware and implementation," *Proceedings of the International Symposium on Robotics Research*, pp. 111 – 116, 1997.

[13] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, UK: Cambridge University Press, 2006.

[14] J. Serra, *Image Analysis and Mathematical Morphology*. Boston: Academic Press, 1982.

[15] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[16] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.